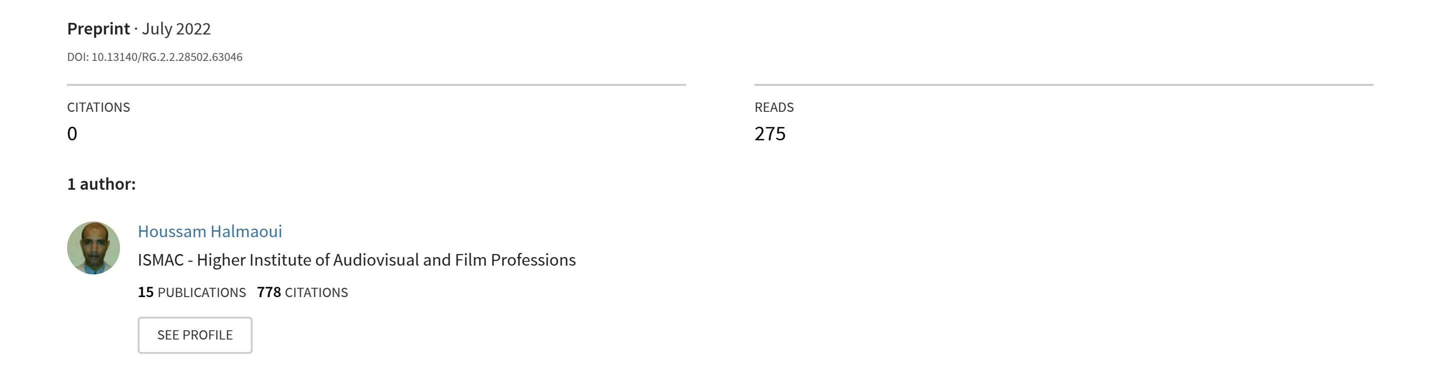
L'intelligence artificielle, la réalité augmentée et la réalité virtuelle dans l'audiovisuel et le cinéma



Preprint · July 2022

DOI: 10.13140/RG.2.2.28502.63046

1 author:

Houssam Halmaoui
ISMAC - Higher Institute of Audiovisual and Film Professions

L’intelligence artificielle, la réalité augmentée et la réalité virtuelle dans l’audiovisuel et le cinéma

Houssam Halmaoui
ISMAC - Institut supérieur des métiers de l’audiovisuel et du cinéma, Rabat, Maroc

Résumé : Avec le développement du cinéma numérique, les algorithmes de vision par ordinateur sont devenus omniprésents et indispensables dans toutes les étapes de fabrication d’un produit audiovisuel. Au Maroc, dans la communauté des chercheurs en informatique, il existe très peu d’applications en audiovisuel à cause de la méconnaissance des besoins des artistes et de l’industrie cinématographique. De l’autre coté, les artistes ont peu de connaissances des dernières avancées en informatique de l’image car ils passent plus de temps sur des tâches de création artistique. Afin de surmonter ce gap de communication, nous présentons ici les techniques récentes de l’intelligence artificielle (IA), de la réalité augmentée (AR : *augmented reality*) et de la réalité virtuelle (VR : *virtual reality*), appliquées à l’audiovisuel. L’un des grands intérêts de l’IA est d’assister les artistes dans le but de leur permettre d’être plus créatifs grâce à l’automatisation et l’accélération de certaines tâches et la qualité de rendu des images. Les applications IA qui ont révolutionné les paradigmes dans le domaine de l’image sont nombreuses : génération de visages et de scènes de synthèse ultra-photoréalistes, animation de portraits et de personnages ou la création de modèles 3D à partir de photos réelles. En ce qui concerne les nouvelles formes narratives, l’écriture et la production pour le web et pour les nouveaux supports numériques nécessitent une connaissance des différentes possibilités offertes par les outils web, les logiciels et les matériels de l’AR et de la VR, qui permettent de créer une interactivité avec le spectateur ou de le faire bénéficier d’une plus grande immersion. Il est aussi important de connaître les limites de ces différentes technologies, tels que la qualité du rendu, les temps de traitement et de travail et l’accessibilité des outils.

1) Introduction

Les années 1990 ont marqué le début de la révolution numérique dans le cinéma et plus particulièrement dans le domaine des effets visuels ou VFX (*Visual Effects*). Cela a été du principalement à l’amélioration des performances des ordinateurs - devenus capables de traiter rapidement les grandes quantités de données dont sont constituées les images - et des autres dispositifs numériques (caméras, projecteurs, etc.), offrant ainsi une meilleure qualité d’image en termes de détails, des couleurs, de rendu 3D, etc. Mais, selon notre observation, le concept de réalisation filmique, en dehors de la qualité offerte par les capteurs et l’électronique numérique, n’a

pas beaucoup changé du point de vue matériel, dans le sens où l'acquisition et la projection des images s'effectuent toujours avec le même type d'instruments. Toutefois, nous assistons aujourd'hui à l'émergence de nouveaux outils de cinéma stéréoscopique, de l'AR et de la VR, mais qui n'ont pas encore réussi à remplacer les écrans traditionnels. Notre constat est que la principale révolution numérique qu'a connue le cinéma ces 30 dernières années est plutôt d'ordre algorithmique, c'est à dire les méthodes calculatoires utilisées par les différentes fonctionnalités logiciels. Dans le cas des VFX, ils permettent, par exemple, d'une manière visuellement parfaite (photoréaliste), de fusionner des images différentes, de supprimer des objets d'une image, de modifier les couleurs d'un objet, de combiner des objets et des environnements 3D de synthèse avec des images réelles, etc. Cela a eu pour conséquence, d'un coté, de modifier complétement le workflow de post-production et de l'autre coté, d'offrir aux réalisateurs de nouveaux outils de fabrication d'images photoréalistes. Aussi, certains algorithmes peuvent être embarqués dans la caméra ou s'exécuter en temps réel pour permettre la pré-visualisation des VFX sur le plateau de tournage.

Nous pouvons catégoriser ces algorithmes en deux types : traditionnels et IA. Les premiers exécutent des tâches selon un enchainement logique et des opérations bien définies implémentées par le programmeur. Les algorithmes d'IA sont fondés sur la notion d'apprentissage à partir des données (image, son, texte, etc.). Bien que l'IA existe depuis le début de l'informatique, le big bang de l'IA n'a débuté qu'en 2012 grâce à l'abondance des données numériques utilisées pour l'apprentissage et à la découverte de nouveaux algorithmes plus performants de *deep learning*, essentiellement dans le domaine de l'image qui est celui où l'IA donne les meilleurs résultats.

Aujourd'hui, malgré l'abondance des outils algorithmiques, avec beaucoup d'applications potentielles dans le domaine de l'audiovisuel, au Maroc, il existe un manque de communication entre informaticiens et artistes. D'un coté, les chercheurs en informatique ne connaissent pas les besoins de l'industrie cinématographique et de l'autre coté, les artistes ont peu de connaissances des nouveaux outils de l'IA, de l'AR et de la VR.

Le but de cet article est de tenter de réduire ce gap de communication en présentant ces différents outils et des exemples d'applications VFX.

Dans la section 2, nous allons présenter une introduction à l'IA et aux notions d'apprentissage automatique, de classification, de génération de données et de photoréalisme. Dans la section 3, nous allons découvrir les différentes applications récentes dans le domaine de l'audiovisuel. La section 4 est consacrée à la découverte des outils de l'AR et de la VR. Enfin, dans la section 5, nous

proposons des solutions pour l'intégration des différentes techniques étudiées dans l'enseignement et la recherche de l'audiovisuel et du cinéma.

2) Intelligence artificielle appliquée à l'image

Avant d'aborder les applications IA dans l'audiovisuel, nous allons commencer par définir les notions fondamentales d'entrainement des modèles IA, de classification et de génération des données.

2.1) Apprentissage automatique et modèles génératifs

L'apprentissage automatique est une branche de l'IA fondée sur les probabilités et permettant à une machine d'apprendre, à partir d'un ensemble de données dit d'apprentissage, à exécuter une tâche spécifique de classification ou de génération de nouvelles données. Il s'agit de la sous-discipline IA la plus utilisée dans les applications de l'audiovisuel.

Afin de comprendre ce concept, nous pouvons faire une analogie avec l'intelligence humaine dans le cas de l'apprentissage des noms des objets par un enfant à qui nous montrons quelques objets d'une certaine classe (chaise, table, téléphone, etc.), en lui indiquant à chaque fois le nom de la classe. Après cette phase d'apprentissage, l'enfant sera capable de reconnaître des nouveaux objets des mêmes types de classes et qu'il n'a jamais vu auparavant. De la même façon, nous fournissons à un modèle IA (programme informatique) des données en grandes quantités étiquetées avec les labels des classes correspondantes, afin de l'entrainer à classifier de nouvelles données similaires. Généralement, plus les données sont nombreuses et diverses, plus le modèle est performant.

À l'inverse des modèles de classification (discriminatoires), nous pouvons créer des modèles de génération des données (génératifs), à qui nous fournissons le nom de la classe (homme, femme, animal, voiture, etc.), afin qu'ils synthétisent des images appartenant à cette classe. Il s'agit des modèles qui ont fait le succès de l'IA dans l'audiovisuel grâce à la synthèse d'images de visages ultra-photoréalistes (*deepfakes*).

L'entrainement de ce type de modèles génératifs peut être expliqué par une analogie : un faussaire d'œuvres d'art essaie de créer des peintures indiscernables des originales et en même temps, un expert tente de différencier les contrefaçons des originales [1]. Les deux protagonistes s'améliorent en apprenant l'un de l'autre. Nous avons donc deux modèles IA : un générateur et un discriminateur. Au début, le générateur ne sait générer que du bruit, que le discriminateur distingue facilement des données réelles. Pour s'améliorer, le générateur observe le résultat de classification et après plusieurs répétitions, il arrive à générer des images de synthèse très photoréalistes qui

peuvent tromper le discriminateur. Après la phase d'apprentissage, le modèle générateur peut être utilisé pour créer de nouvelles données de synthèse de manière automatique.

L'un des grands intérêts des modèles génératifs est, qu'à la différence des méthodes traditionnelles de synthèse d'images à l'aide de logiciels de modélisation 3D, qui nécessitent un important temps d'apprentissage des outils logiciels (les plus compliqués dans l'audiovisuel), les modèles IA permettent de générer du contenu de manière automatique et rapide, ce qui laisse le temps aux artistes de passer plus de temps sur la partie créative.

Il existe deux grands challenges aujourd'hui : le photoréalisme et le contrôle du contenu généré, que nous allons détailler dans ce qui suit.

2.2) Photoréalisme

La qualité des VFX se mesure en fonction de leur niveau de "photoréalisme" qui donne au film plus de vraisemblance et de crédibilité. Nous parlons de "photoréalisme" plutôt que de "réalisme" pour distinguer les éléments créés sur ordinateur des objets réels.

Dans le cas des modèles IA, afin de générer des images photoréalistes indiscernables des photographies, il faut à la fois utiliser de très larges ensembles de données d'apprentissage (des millions d'images), des super-ordinateurs capables de traiter de telles quantités de données et des algorithmes (architecture des modèles) performants. La collecte des données est l'étape qui demande le plus de temps de travail. Certains modèles entrainés avec des données de synthèse (générés par d'autres modèles et donc plus faciles à collecter) donnent de très bons résultats [2]. Les ordinateurs dédiés à l'IA sont de plus en plus accessibles et de même pour les outils logiciels de programmation gratuits qui sont supportés par une grande communauté. Concernant les algorithmes, depuis le développement de l'apprentissage profond (*deep learning*) - qui consiste en l'entrainement de modèles complexes à partir de très large quantité de données - lors de la dernière décennie, les articles de recherche sont devenus abondants et la difficulté réside plutôt dans le choix de la méthode la plus adéquate parmi des centaines et son adaptation pour un problème spécifique.

La figure 1 montre des images obtenues par un modèle simple que nous avons entrainé à l'aide de données constituées de 1000 images de visages de célébrités en basse résolution (issues de la base d'images de [3]) sur un ordinateur de moyenne gamme. Le temps d'apprentissage est de quelques heures, mais la génération est instantanée. À l'opposé, nous montrons le résultat très photoréaliste obtenu par un modèle plus complexe [4] entrainé pendant plusieurs jours sur une plus large base d'images haute résolution à l'aide d'un ordinateur performant (calculs effectués sur une carte graphique puissante).



Figure 1 : Images générées par apprentissage simple (à gauche) et profond (à droite).

2.3) Contrôle du contenu : 3D traditionnelle et rendu neuronal

Nous pouvons classer les modèles génératifs principalement en deux types :

- Conditionnels : synthèse d'objets d'une classe spécifique de manière aléatoire.

- Contrôlables : spécification d'une ou de plusieurs caractéristiques dans le contenu de l'image.

Le challenge des modèles génératifs contrôlables est d'avoir, d'une manière similaire à la synthèse des images par des logiciels de modélisation et d'animation 3D, un grand degré de contrôle sur le contenu des images (cheveux, yeux, lunettes, etc.), ainsi que sur la position, l'orientation et l'animation des objets et l'éclairage de la scène.

En effet, la pipeline classique de la 3D est la suivante : modélisation, éclairage de la scène, animation et rendu, avec des méthodes utilisant des formules de la physique pour simuler l'apparence des matériaux, l'éclairage et la mécanique des objets de la scène. Comme nous l'avons signalé précédemment, chaque étape de cette pipeline nécessite l'apprentissage des outils logiciels correspondants. Dans le cinéma d'animation, des équipes de plusieurs personnes travaillent chacune sur un seul aspect, ce qui est couteux en temps et en argent.

Récemment, une nouvelle discipline de l'IA, appelée rendu neuronal combinant à la fois les méthodes d'apprentissage automatique et l'aspect physique de l'infographie 3D, permet d'automatiser la synthèse, le rendu et l'animation des objets et de créer des scènes plus photoréalistes qu'avec les méthodes traditionnelles. Dans [5], nous avons effectué une étude détaillée des méthodes par IA et une comparaison avec la 3D traditionnelle. Dans la section

suivante, nous allons présenter différents exemples de modèles génératifs contrôlables et des applications existantes et potentielles dans l'audiovisuel.

3) Applications dans l'audiovisuel

Les exemples suivants montrent comment l'IA peut améliorer le workflow dans l'audiovisuel et les VFX, grâce à l'automatisation et l'accélération de certaines tâches - qui nécessitent des temps de travail importants avec les logiciels traditionnels de post-production - ce qui a l'avantage de laisser aux artistes plus de temps pour exprimer leur créativité.

3.1) Synthèse de vues nouvelles

La synthèse de vues nouvelles consiste à effectuer, à partir d'une ou de plusieurs images acquises avec différentes positions caméra, le rendu photoréaliste d'une image de la scène acquise d'une quelconque position caméra [6]. Certaines méthodes permettent de générer un modèle 3D à partir d'une seule image [7]. Dans [8], un modèle permet de générer des vidéos photoréalistes.

Nous pouvons imaginer une application audiovisuelle de simulation de plans multi-caméras en utilisant un nombre réduit de caméras sur le plateau. Ceci permettra au monteur et au réalisateur d'avoir plus de matière pour travailler et donc d'être plus créatifs.

3.2) Le ré-éclairage

Le ré-éclairage consiste à générer, à partir d'images acquises d'un même point de vue mais avec différentes positions d'éclairage, une nouvelle image de la scène en simulant une position d'éclairage quelconque. Dans [9], un modèle permet cela en utilisant seulement 5 images différentes (5 positions d'éclairage), alors que d'autres modèles [10] nécessitent une dizaine ou une centaine d'images mais permettent un rendu plus photoréaliste. Certaines méthodes récentes utilisent seulement une seule position d'éclairage [11] . De nouvelles méthodes permettent le ré-éclairage d'objets spécifiques : portrait [12], corps[13] et scène d'extérieur [14].

Le ré-éclairage peut avoir un rôle d'une importance cruciale dans les applications audiovisuelles pour corriger des problèmes techniques ou esthétiques d'éclairage. Par exemple, lors du tournage d'une scène, l'éclairage peut ne pas correspondre à celui souhaité par le réalisateur, soit à cause d'un changement imprévisible dans la météo, un décalage horaire dans le planning du tournage ou une mauvaise position d'éclairage. Dans de tels cas, pour éviter de refaire le tournage ou de passer beaucoup de temps en post-production pour corriger le problème, nous pouvons utiliser un modèle de ré-éclairage en lui fournissant uniquement la position de l'éclairage souhaité. Le résultat sera

automatique et rapide et donc beaucoup plus économique. Aussi, de la même façon que l’étalonnage des couleurs, le ré-éclairage peut être utilisé à des fins créatives pour donner un certain aspect artistique aux images. Notons enfin qu’en combinaison avec la synthèse de vues nouvelles, le ré-éclairage permettra de contrôler à la fois l'éclairage et la position de la caméra.

3.3) Contrôle de contenu et animation

Concernant l’animation, la plupart des travaux existants ont été effectués sur des images de visages. Dans [15], le modèle proposé permet de transférer d’une vidéo source à une vidéo cible, à la fois la position de la tête, les expressions faciales et la direction du regard. Dans [16], un modèle similaire permet de transférer les mêmes paramètres, mais à partir d’un simple dessin de visage fourni par l’utilisateur.

Plusieurs applications existent déjà dans le cinéma ou la télévision. Nous citons l’émission récente *Hôtel du temps* (France 3, 2022) dont l’objectif est de faire des interviews avec des célébrités décédées en reconstituant leurs visages et leurs voix par apprentissage sur des données image et audio provenant des archives. Les résultats visuels sont ultra-photoréalistes. Concernant le clonage de voix, une autre application potentielle est le doublage avec les vraies voix des acteurs.

L’un des avantages de ces modèles pour les productions audiovisuelles est de faire jouer des acteurs sans qu’ils ne soient présents sur le plateau de tournage et donc de faire des économies par rapport au cas d’une présence réelle de l’acteur, avec un résultat similaire, voire meilleur, étant donnée la possibilité de contrôler les expressions faciales et l’apparence des acteurs d’une manière précise.

3.4) Traduction texte en image

Certains modèles sont capables de générer des images à partir d’un texte décrivant le contenu souhaité. Les résultats obtenus par les modèles les plus récents tel que Dall-E 2 [17] sont d’un niveau de photoréalisme et d’une qualité de détails sans précédent. D’autres modèles, permettent plutôt de générer du contenu d’un certain style de peinture choisi par l’utilisateur [18].

De tels outils peuvent être utilisés pour inspirer et assister les artistes en transformant leurs pensées en images. Notons qu’avec le même texte, certains IA peuvent générer une image différente à chaque fois et offrir donc à l’artiste plusieurs choix possibles.

L’artiste peut aussi manipuler le contenu en modifiant le vocabulaire utilisé dans le texte. Les mots les plus simples sont traduits plus efficacement en image car ils ont plus de chance d’être présents dans la base d’apprentissage.

Les travaux futurs tentent de générer à la volée des vidéos à partir de texte [19], les résultats sont prometteurs même s'ils sont encore loin de la qualité des scènes réelles.

3.5) Synthèse sémantique d'images

Les modèles de synthèse sémantique d'images [20] offrent la possibilité de générer des images photoréalistes à partir d'un simple croquis des différents objets de la scène. L'utilisateur peut choisir entre différents types d'objets (arbre, ciel, montagne, bâtiment, etc.). De la même façon que pour la traduction de texte en image, cela permettra d'inspirer les artistes ou d'avoir un point de départ pour la création de nouveaux environnements photoréalistes.

3.6) Transfert de style

Il s'agit de modèles permettant d'utiliser le style d'une image et de le transférer à une autre image sans en modifier le contenu. Les exemples d'utilisation peuvent aller de l'étalonnage des couleurs, pour donner un certain aspect artistique aux images, aux VFX afin de créer de nouveaux environnements. Les modèles récents [21] permettent de générer des images ultra-photoréalistes.

3.7) Animation de visages à partir de source audio

Certains modèles [22] permettent d'animer les expressions faciales ou d'effectuer le *Lip-Sync* d'un personnage 3D à partir d'une source audio. Ceci permet de simplifier le workflow d'animation 3D en fournissant aux animateurs un premier résultat à utiliser comme point de départ. Le challenge aujourd'hui est d'inclure aussi les émotions à partir de la voix.

3.8) Super-résolution par IA

Concernant l'augmentation de la résolution des images, les algorithmes IA ont atteint un niveau sans précédent dans la qualité des détails. Des informations spatiales (pixels) non présentes dans l'image originale peuvent être recréées d'une manière lisse, nette et sans introduction des artefacts [23]. Les applications sont d'une importance cruciale pour la restauration et le transfert de la culture audiovisuelle.

3.9) Slow Motion par IA

De la même façon que les méthodes de super-résolution permettent de créer de nouveaux pixels inexistants dans l'images originale, les méthodes de *slow motion* (ralenti) créent de nouvelles images inexistantes dans des vidéos par interpolation entre deux images successives. Les méthodes

récentes d’interpolation par des modèles IA [24] surpassent les algorithmes traditionnels en terme de qualité de détails des objets en mouvement.

3.10) Estimation de la pose humaine

Le but de la détection de la pose humaine est de trouver dans une vidéo la position des différentes articulations du corps afin de créer une représentation du squelette. Pour effectuer la détection il faut prendre en considération plusieurs problèmes : différentes positions, éclairages, formes, dimensions, bruits, etc. Dans le domaine des VFX, la technique la plus utilisée est la capture de mouvement avec de multiples caméras (entre 10 et 20) minutieusement calibrées. Les acteurs portent des costumes spécifiques contenant des marqueurs. Les caméras suivent les marqueurs pour reconstruire le mouvement. Cette approche permet d’obtenir un résultat très précis et avec une haute cadence d’images, mais nécessite un matériel cher, encombrant et calibré. Elle est souvent réservé à des films à grand budget. Les approches IA [25] permettent d’obtenir des résultats rapides et précis avec une seule caméra et dans des environnements non contrôlés. Les applications audiovisuelles vont de la capture de mouvement (remplacer les acteurs par des personnages 3D), au suivi des déplacements des joueurs dans les compétitions sportives.

4) Réalité augmentée et réalité virtuelle

Les dispositifs de création cinématographique ont beaucoup évolué depuis l’apparition du cinéma digital. Toutefois, la projection des scènes filmées par une caméra (ou synthétisées sur ordinateur) sur un écran reste le modèle le plus dominant.

Plusieurs tentatives de changement de cette norme ont eu lieu, depuis les années 1950 (Hitchcock), avec le cinéma stéréoscopique 3D qui n’a pas cessé d’évoluer en terme de qualité de rendu et de perception des profondeurs. Toutefois, elle n’a toujours pas eu le succès tant attendu. En effet, ils sont rares les réalisateurs qui osent expérimenter cette technique pour essayer de raconter une histoire de façon différente et susciter d’autres sensations chez le spectateur par le biais de l’effet d’immersion.

Durant la dernière décennie, de nouvelles formes narratives interactives ont aussi vu le jour sur les plateformes web. L’exemple le plus connu est le web documentaire. De telles productions audiovisuelles permettent au spectateur d’interagir avec l’histoire en se mettant à la place du personnage pour décider du déroulement de l’histoire à des moments précis de la vidéo (série "black mirror") ou en regardant le documentaire de manière non linéaire ("bear 71" ou "la zone") en sélectionnant les séquences à visionner. De même, ce format est très peu exploré par les réalisateurs

à cause de la nécessité d'utilisation des outils de développement web qui sont peu connus dans le domaine de l'audiovisuel. Les scénaristes peuvent aussi s'inspirer des techniques d'écriture pour les jeux vidéos qui partagent beaucoup de points en communs avec les films interactifs.

Aujourd'hui, le matériel et les outils AR et VR sont devenus plus accessibles aux consommateurs et ont ouvert la porte à de nouvelles possibilités de créations audiovisuelles. Dans la suite, nous allons présenter les outils et les techniques de chacune de ces technologies.

4.1) Réalité augmentée

Les VFX, en dehors de leur utilisation pour créer des effets spectaculaires, sont utilisés aujourd'hui dans la plupart des films pour ajouter et supprimer des éléments de l'image (environnement, objet ou personnage), de la même façon que les décors sont construits sur le plateau, mais d'une manière numérique.

L'AR partage plusieurs techniques avec les VFX. Elle consiste en la superposition d'éléments digitaux (images, objets 3D, etc.) à des vidéos réelles. Les éléments virtuels ajoutés doivent coexister de manière cohérente avec les perspectives et l'éclairage de la scène réelle lorsque la caméra se déplace. Pour cela, il est important d'effectuer le suivi du mouvement caméra de manière précise. Nous distinguons trois types d'approches :

- Capteurs : GPS, gyroscope, accéléromètre, etc.
- Vision : utilisation des algorithmes de vision par ordinateur ou de l'IA.
- Hybride : Combinaison des approches capteurs et vision.

Nous avons effectué plusieurs travaux sur les systèmes de vision car ils ne nécessitent aucun matériel supplémentaire et sont donc plus accessibles. Dans [26], nous avons proposé des systèmes de suivi de mouvement pour l'insertion des éléments 2D ou 3D dans des images. Dans [27] et [28], nous avons proposé de nouveaux algorithmes fondés sur l'apprentissage. Dans [29], nous avons effectué une revue et une comparaison de l'état de l'art des méthodes traditionnelles et par IA. Dans [30], nous avons proposé un système de AR en 3D utilisant des marqueurs artificiels, comme le montre la figure 2, où nous avons pris en compte aussi, en plus du suivi de mouvement, l'aspect éclairage des objets 3D insérés. Dans [5], nous avons effectué une revue et une comparaison de l'état de l'art des techniques d'éclairage et de rendu 3D traditionnelles et par IA.

L'objectif principal de nos recherches est de proposer des systèmes AR par des approches de vision par ordinateur afin de permettre la prévisualisation des VFX en temps réel.

Figure 2 : Suivi caméra, insertion et éclairage d'un objet 3D dans une vidéo

Il est aussi possible, par combinaison avec les approches capteurs et les outils VR (et avec la créativité des réalisateurs), de proposer aux spectateurs de nouvelles expériences immersives. Pour ce faire, il est important que les réalisateurs connaissent les différentes possibilités offerte par les outils VR, que nous allons découvrir dans la suite.

4.2) Réalité virtuelle

La définition de la VR change avec le temps. Plus récemment, le terme VR est souvent utilisé pour désigner exclusivement les expériences immersives et interactives visualisées grâce à des appareils VR spécifiques. Les casques HMD (Head Mounted Display) sont les appareils les plus utilisés, mais il existe aussi des systèmes moins connus du public comme CAVE (Cave Automatic Virtual Environnement) où l'image est projetée sur différents murs d'une pièce.

Trois choses rendent la VR plus immersive que n'importe quel autre média (télévision, cinéma, TV3D et cinéma 3D) :

- Stéréo-vision 3D.

- Contrôle dynamique du point de vue l'utilisateur.

- Expérience environnante (*surround*).

Nous allons comparer les systèmes HMD et CAVE selon ces trois critères.

Casques HMD

Le casque est constitué de deux afficheurs (un pour chaque œil). Les deux images sont légèrement différentes (comme pour la vision humaine), afin de permettre une vision stéréo. Le principe est similaire au cinéma 3D mais l'expérience est plus réelle. En effet, au cinéma, plusieurs spectateurs regardent un seul écran, alors que dans le cas de la VR le point de vue est adapté à la position exacte

d'un utilisateur spécifique. Ceci est réalisé grâce à un appareil de suivi des mouvements de la tête, intégré au casque.

D'un autre coté, la perception visuelle du spectateur est complétement entourée par le casque. En général, notre expérience diffère selon si nous regardons un film sur la télévision ou dans une salle de cinéma : plus notre champs de vision est couvert par l'écran, plus l'expérience est immersive. La taille des casques VR est plutôt petite, mais rien ne couvre notre champs de vision à part les images (il n'y a pas d'échappatoire), et à la différence du cinéma 3D, le spectateur continue de regarder l'image virtuelle même s'il tourne la tête.

Système CAVE

Le CAVE est une pièce vide constituée d'au moins 4 murs d'affichage (3 murs aux alentours et le sol). Des projecteurs d'image en haute résolution sont placés derrière les murs. L'utilisateur porte des lunettes 3D pour avoir une vision stéréo. Même si l'utilisateur est entouré de murs réels, il a l'impression d'être dans un environnement virtuel. Le point de vue est contrôlé de manière similaire au HMD grâce au suivi de mouvement de la tête par des capteurs. Mais, même s'il est possible d'avoir plusieurs utilisateurs dans le même espace, ayant chacun une paire de lunette, un seul capteur de mouvement peut être synchronisé avec l'ordinateur qui effectue le rendu des images projetées (la vision des autres utilisateurs sera déformée). Il est possible d'avoir des expériences VR multi-utilisateurs, mais il faut un système d'affichage propre à chaque utilisateur.

En comparaison, le système CAVE possède un grand écran avec une résolution très supérieure à celle du HMD, mais il nécessite un espace fixe et dédiée alors que le HMD est nomade et moins encombrant.

Capteurs

Le suivi du point de vue doit être réalisé dans l'espace 3D : rotation et position de la tête.

Le suivi de la rotation de la tête est réalisé grâce à un accéléromètre, un gyromètre ou les deux. De tels systèmes sont utilisés dans les smartphones pour détecter sa rotation.

Le suivi de la position de la tête est réalisé grâce à des caméras externes. Le système CAVE utilise plusieurs caméras infrarouges. Les HMD utilisent une seule caméra infrarouge placé devant l'utilisateur. Le suivi échoue si l'utilisateur se trouve en dehors du champ de vision de la caméra.

Certains appareils VR ne disposent que du suivi de rotation, ce qui oblige l'utilisateur à rester immobile et à ne pas se déplacer comme il le souhaite, mais uniquement tourner sa tête. Il lui est donc impossible de se rapprocher des objets ou de les voir de différents points de vue.

Notons qu’il est toujours possible de naviguer dans l’environnement 3D avec un contrôleur (que nous allons présenter dans la suite), mais cela ne sera pas aussi naturel que dans la vie réelle.

Généralement, seuls les appareils VR haut de gamme disposent d’un capteur de position qui nécessite une certaine puissance de calcul pour effectuer le suivi de manière robuste.

Contrôleur et dispositif haptique

Les contrôleurs permettent d’interagir avec l’environnement virtuel ou de communiquer par des gestes. Pour cela, le système VR a besoin de connaître la position et la rotation des mains, qui sont calculés, comme pour le suivi de la tête, à l’aide de capteurs de suivi intégrés au contrôleur pour la rotation et externes pour la position.

Le contrôleur permet aussi quelque chose que nous ne pouvons pas faire avec nos mains dans la vraie vie : la navigation dans l’environnement. L’intérêt est d’explorer un monde virtuel plus grand que l’espace réel où se trouve l’utilisateur. Notons qu’il y a un risque de ressentir des nausées et des vertiges causés par l’utilisation des manettes du contrôleur, cet effet est plus important avec le HMD qu’avec le CAVE car l’utilisateur bouge moins son corps.

La plupart des contrôleurs sont capables aujourd’hui de produire un retour de force par vibration (*Haptic Feedback*). En effet, comme le système connaît la position des objets présents dans l’environnement 3D, ainsi que celle de l’utilisateur, lorsque les deux interagissent ensemble, il est possible de générer une vibration similaire à ce que nous ressentons dans la vie réelle. Cela permet une plus forte sensation d’immersion. Mais parfois l’effet est tellement convaincant que l’utilisateur peut oublier que les objets virtuels n’existent pas et interagir avec les vrais objets se trouvant à coté (qu’il ne voit pas), ce qui peut causer des accidents.

HMD : mobile, console et haute gamme

Les caques HMD mobiles sont les plus accessibles. Ils nécessitent uniquement un téléphone mobile mais ne permettent qu’un certain suivi de rotation. De l’autre coté, nous avons des consoles VR dédiées avec plus de fonctionnalités grâce au contrôleur et à la caméra de suivi de position. Enfin, pour une meilleure expérience, il faut utiliser un ordinateur haut de gamme avec une carte graphique performante. Les appareils mobiles sont les plus disponibles et reste donc la meilleure solution pour cibler le grand public.

Pour montrer une création audiovisuelle dans une exposition artistique, il faut opter pour la solution haut de gamme qui va permettre d’obtenir le niveau d’immersion nécessaire pour exprimer tout le potentiel de l’idée du projet.

Enfin, notons que l'utilisation de la VR pour des créations audiovisuelles nécessite la collaboration avec des acteurs de différentes spécialités pour la création de l'environnement virtuel 3D et pour le développement informatique de l'aspect interactif.

5) Intégration dans l'enseignement et la recherche

Avec l'émergence des nouveaux outils IA, AR et VR, nous assistons aujourd'hui à une phase transitoire comparable à celle de la révolution numérique dans le cinéma au début des années 1990. En effet, les premiers algorithmes VFX ont été empruntés des domaines de la recherche scientifique et militaire et adaptés au cinéma, à travers la collaboration des artistes avec des chercheurs et des informaticiens [31]. Avant l'apparition des logiciels destinés au grand public, seuls les grands studios possédaient les algorithmes nécessaires qu'ils créaient eux mêmes par collaboration avec des acteurs du monde académique et industriel.

Aujourd'hui encore, les techniques IA présentées dans cet article ne sont toujours pas accessibles dans les logiciels VFX disponibles, mais plusieurs réalisateurs dans le cinéma ou la télévision ont pu les utiliser dans leurs projets en collaborant avec des informaticiens.

Pour accompagner cette transition dans l'enseignement et la recherche dans le cinéma, il est fondamental de repenser les métiers de l'audiovisuel en fonction de l'évolution des nouvelles technologies. Pour cela, il est nécessaire de sensibiliser les étudiants et les artistes à l'IA, l'AR, la VR et aux outils web. En même temps, il est essentiel d'expérimenter leur utilisation dans des projets audiovisuels et filmiques, à l'école et en milieu professionnel, en collaborant avec des spécialistes en informatique. Il est aussi important de donner naissance à des projets de partenariat pluri-disciplinaires entre des chercheurs dans le cinéma et dans l'informatique, des artistes et des industriels, dans le but de comprendre et de trouver des solutions innovantes aux besoins de l'industrie cinématographique.

6) Conclusion

Les algorithmes VFX jouent un rôle important dans le développement du cinéma en contribuant à améliorer la narration visuelle, à pallier les limitations du tournage et à réduire les coûts. Les VFX fondés sur l'IA surpassent aujourd'hui les techniques traditionelles par leur qualité de rendu ultra-photoréaliste. Ils permettent aussi d'améliorer le workflow des artistes grâce à la génération de contenu image de manière contrôlable et automatique. Tout cela, permet de réduire considérablement les temps de travail et d'apprentissage des outils par l'utilisateur. Les applications assez récentes, existantes et potentielles dans l'audiovisuel, témoignent de l'efficacité de ces

algorithmes. De l'autre coté, les outils de l'AR et de la VR offrent aux artistes de nouveaux moyens de prévisualisation des VFX en temps réel et de création audiovisuelle permettant des niveaux d'immersion plus grands que tous les médias existants. Le matériel de l'AR et de la VR est en passe de devenir de plus en plus accessible au grand public. Les algorithmes VFX fondés sur l'IA sont assez récents et ne sont pas encore implémentés dans les logiciels existants. En plus de ces questions d'accessibilité, les artistes connaissent très peu ces outils. Pour toutes ces raisons, des collaborations pluri-disciplinaires entre les différents acteurs du monde académique, les artistes et les spécialistes de l'IA sont donc nécessaires. L'accompagnement de cette transition technologique passe aussi par la sensibilisation des étudiants, l'expérimentation et un travail de réflexion sur l'avenir des métiers de l'audiovisuel.

Références

[1] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

[2] Jahanian, Ali, et al. "Generative models as a data source for multiview representation learning." *arXiv preprint arXiv:2106.05258* (2021).

[3] Liu, Ziwei, et al. "Deep learning face attributes in the wild." *Proceedings of the IEEE international conference on computer vision*. 2015.

[4] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[5] Halmaoui, H., Haqiq, A.: Computer graphics rendering survey: From rasterization and ray tracing to deep learning. In: International Conference on Innovations in Bio-Inspired Computing and Applications. pp. 537–548.Springer (2021)

[6] Eslami, SM Ali, et al. "Neural scene representation and rendering." *Science* 360.6394 (2018): 1204-1210.

[7] Wiles, Olivia, et al. "Synsin: End-to-end view synthesis from a single image." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[8] Lombardi, Stephen, et al. "Neural volumes: Learning dynamic renderable volumes from images." *arXiv preprint arXiv:1906.07751* (2019).

[9] Xu, Zexiang, et al. "Deep image-based relighting from optimal sparse samples." *ACM Transactions on Graphics (ToG)* 37.4 (2018): 1-13.

[10] Ren, Peiran, et al. "Image based relighting using neural networks." *ACM Transactions on Graphics (ToG)* 34.4 (2015): 1-12

[11] Griffiths, David, Tobias Ritschel, and Julien Philip. "OutCast: Outdoor Single-image Relighting with Cast Shadows." *Computer Graphics Forum*. Vol. 41. No. 2. 2022.

[12] Sun, Tiancheng, et al. "Single image portrait relighting." *ACM Trans. Graph.* 38.4 (2019): 79-1.

[13] Kanamori, Yoshihiro, and Yuki Endo. "Relighting humans: occlusion-aware inverse rendering for full-body human images." *arXiv preprint arXiv:1908.02714* (2019).

[14] Philip, Julien, et al. "Multi-view relighting using a geometry-aware network." *ACM Trans. Graph.* 38.4 (2019): 78-1.

[15] Kim, Hyeongwoo, et al. "Deep video portraits." *ACM Transactions on Graphics (TOG)* 37.4 (2018): 1-14.

[16] Zakharov, Egor, et al. "Few-shot adversarial learning of realistic neural talking head models." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[17] Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." *arXiv preprint arXiv:2204.06125* (2022).

[18] Crowson, Katherine, et al. "Vqgan-clip: Open domain image generation and editing with natural language guidance." *arXiv preprint arXiv:2204.08583* (2022).

[19] Li, Yitong, et al. "Video generation from text." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.

[20] Park, Taesung, et al. "GauGAN: semantic image synthesis with spatially adaptive normalization." *ACM SIGGRAPH 2019 Real-Time Live!*. 2019. 1-1.

[21] Li, Yijun, et al. "A closed-form solution to photorealistic image stylization." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[22] Karras, Tero, et al. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." *ACM Transactions on Graphics (TOG)* 36.4 (2017): 1-12.

[23] Wang, Xintao, et al. "Real-esrgan: Training real-world blind super-resolution with pure synthetic data." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[24] Jiang, Huaizu, et al. "Super slomo: High quality estimation of multiple intermediate frames for video interpolation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[25] Rafi, Umer, et al. "An Efficient Convolutional Network for Human Pose Estimation." *BMVC*. Vol. 1. 2016.

[26] Halmaoui, H., Haqiq, A.: Feature detection and tracking for visual effects: Augmented reality and video stabilization. In: International Conference on Artificial Intelligence & Industrial Applications. pp. 291–311. Springer (2020)
[27] Halmaoui, Houssam, and Abdelkrim Haqiq. "Convolutional sliding window based model and synthetic dataset for fast feature detection." *The International Conference on Artificial Intelligence and Computer Vision*. Springer, Cham, 2021.
[28] Halmaoui, Houssam, and Abdelkrim Haqiq. "Synthetic feature pairs dataset and siamese convolutional model for image matching." *Data in Brief* 41 (2022): 107965.
[29] Halmaoui, Houssam, and Abdelkrim Haqiq. "Feature matching for 3D AR: Review from handcrafted methods to deep learning." *International Journal of Hybrid Intelligent Systems* Preprint (2021): 1-20.
[30] Halmaoui, H., Haqiq, A.: Matchmoving previsualization based on artificial marker detection. In: International Conference on Advanced Intelligent Systems and Informatics. pp. 79–89. Springer (2020)
[31] Venkatasawmy, Rama. *The Digitization of Cinematic Visual Effects: Hollywood's Coming of Age*. Lexington Books, 2012.